Building a bilingual dictionary from movie subtitles based on inter-lingual triggers

Caroline Lavecchia, Kamel Smaïli, David Langlois

Building a bilingual dictionary from movie subtitles based on inter-lingual triggers

Lavecchia Caroline[1], Smaïli Kamel[1], and Langlois David[12]

[1] LORIA, Campus Scientifique, BP239, 54506 Vandoeuvre-lès-Nancy, FRANCE
[2] IUFM of Lorraine
{lavecchi,smaili,langlois}@loria.fr

Abstract. This paper focuses on two aspects of Machine Translation: parallel corpora and translation model. First, we present a method to automatically build parallel corpora from subtitle files. We use subtitle files gathered from the Internet. This leads to useful data for Subtitling Machine Translation. Our method is based on Dynamic Time Warping. We evaluated this alignment method by comparing it with a sample aligned by hand and we obtained a precision of alignment equal to 0.92. Second, we use the notion of inter-lingual triggers in order to build from the subtitle parallel corpora multilingual dictionaries and translation tables for machine translation. Inter-lingual triggers allow to detect couple of source and target words from parallel corpora. The Mutual Information measure used to determine inter-lingual triggers allows to hypothesize that a word in the source language is a translation of another word in the target language. We evaluate the obtained dictionary by comparing it to two existing dictionaries. Then, we integrated the obtained translation tables into an entire translation decoding process supplied by Pharaoh (Koehn, 2004). We compared the translation performance using our translation tables with the performance obtained by the Giza++ tool (Al-Onaizan et al., 1999). The results showed that the system tuned for our tables improves the Bleu (Papineni and al., 2001) value by 2.2% compared to the ones obtained by Giza++.

1 Introduction

Training machine translation systems requires a huge quantity of bilingual aligned corpora. Even if this kind of corpora becomes increasingly available, there may be a coverage problem for a specific need. Building bilingual parallel corpora is an important issue in machine translation. Several French-English applications use either the Canadian Hansard corpus or corpora extracted from the proceedings of European Parliament (Koehn, 2005). One way to enrich the existing parallel corpora is to catch the important amount of free available movie subtitles. Several web-sites (*http://divxsubtitles.net*) provide files used for subtitling movies. This quantity of information may enhance the existing bilingual corpora and enlarges the nowadays-covered areas. Furthermore, subtitles corpora are very attractive due to the used spontaneous language which contains formal, informal and in some movies vulgar words. Consequently, working on parallel movie corpora constitutes a good challenge to go towards realistic translation machine applications. Movies corpora include so many common expressions, hesitations, coarse

words,... Training translation models on these corpora will lead to spontaneous speech translation machine systems dedicated to a large community. This work could be considered as a first stage towards a real time subtitling machine translation system.
For one movie, two subtitle files for two different languages are not necessary aligned because the different files are independently made by different human translators. The raw subtitle corpora cannot be used without pre-processing. In order to make these files convenient for use, it is first necessary to align bilingual versions of the same movie at paragraph, sentence or phrase level. The raw data are difficult to align because subtitles are segmented such that it is easy to read and to write them on screen. Therefore, a sentence may be segmented into several phrases: usually, subtitles are presented on two lines of 32 characters which are readable on six seconds in maximum (Vandeghinste and Sang, 2004). Moreover, the constraint of segmentation applies differently from one language to another because of the language features.

One of the objectives of this paper is to present a method which automatically aligns two subtitle files. This method is based on Dynamic Time Warping (DTW) algorithm. In the following, we pinpoint the specific features of subtitles and present a measure which helps to align them efficiently.
The second objective of this paper is to use this parallel corpus to train the parameters of a machine translation system. In this scope, we propose an original method to construct automatically a translation table without any external knowledge. This objective is handled by inter-lingual triggers which are used to induce a bilingual dictionary – which overcomes the need of building up a dictionary by hand – and the parameters of the translation table. We describe the idea of inter-lingual triggers, the way to exploit and to make good use of them in order to produce a bilingual dictionary. Then, we describe how to compute the corresponding translation table. Finally, our translation table is evaluated by comparing it to the one achieved by Giza++ (Al-Onaizan et al., 1999) in an entire translation decoding process supplied by Pharaoh (Koehn, 2004). The experiments show that the obtained translation table is well constructed and is suitable for machine translation. In a tuned use of Pharaoh parameters, our model can outperform the model 3 of Giza++.

2 Building parallel corpora from movie subtitles

Our objective is to obtain as much pairs of aligned subtitles from movie subtitles as possible. Two subtitles are aligned if they are translations of one another.

2.1 Data description

A subtitle file is a set of phrases or words corresponding to: a set of dialogues, a description of an event or a translation of strings displayed on screen (in general destinated to deaf people, or people without skills in the original language). A subtitle is a textual data usually presented at the bottom of the screen. The text could be written in original version or in a foreign language and corresponds to what is being said by an actor

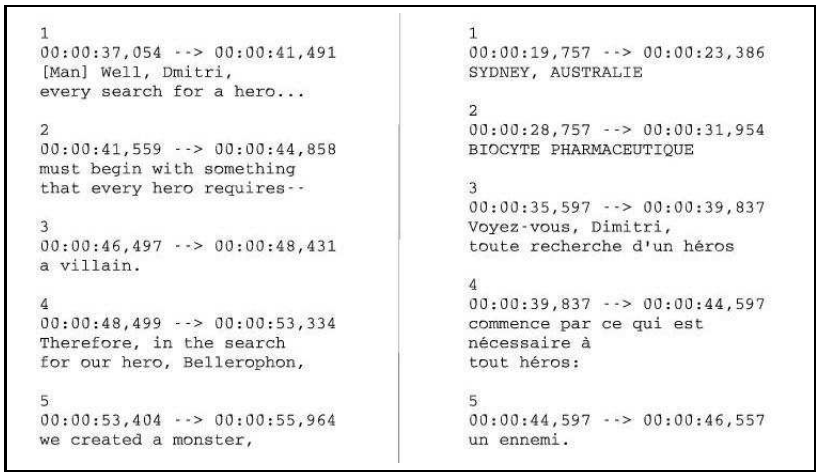

Fig. 1: Source and target movie subtitles

or what is being described. Fig. 1 shows a piece of subtitles extracted from the movie *Mission Impossible 2*.

Each subtitle is characterized by an identifier, a time frame and a sequence of words. The time frame indicates the interval time the subtitle becomes visible on the screen. The sequence of words corresponds to the literal version of the dialogue or to the reported event. Subtitles as they are presented can not be used directly for alignment because the French and English subtitles do not match. In the example of Fig.1, the content of the first two subtitles mismatch. In fact the English subtitle begins with a dialogue when the French one does not. Because the movie is American, if any informative message is displayed on the screen, it is thus not necessary to repeat it into the English subtitle file. In the opposite in French the translation is necessary. This kind of difference occurs very frequently and produces gaps between the French and the English subtitles. Several other errors can be present in the subtitle files: omission, insertion of subtitles,... For more details one can refer to (Lavecchia et al., 2007).

2.2 Alignment solutions

The major works aiming at solving the alignment of parallel corpora are based on dynamic programming. These works use a distance measure to evaluate the closeness between corpus segments. A segment can be a paragraph, a sentence or a phrase. The segmentation may be available or calculated automatically as in (Melamed, 1996). Several solutions and different options have been proposed, for more details we can refer to (Moore, 2002; Brown et al., 1991; Melamed, 1996; Vandeghinste and Sang, 2004; Gale and Church, 1991). One can find a comparative study about several methods in (Singh and Husain, 2005).

2.3 Dynamic Time Warping based on F-measure

Aligning two subtitle files can be considered as a classical problem of dynamic programming. As shown previously, English and French subtitles are asynchronous. To align them, we utilize DTW based on F-measure. This measure is used to calculate the best path between two subtitle files. Intuitively, two subtitles are not aligned if none or only few words of source and target match. This leads to hypothesize that two subtitles do not match if their F-measure is weak.
In Fig. 2, each node (e, f) represents a potential matching point between one English and one French subtitle. A correct path begins at node $(0,0)$ and ends at node (E,F) where E is the number of English subtitles and F the number of French subtitles. From a node, the following shifts are possible:

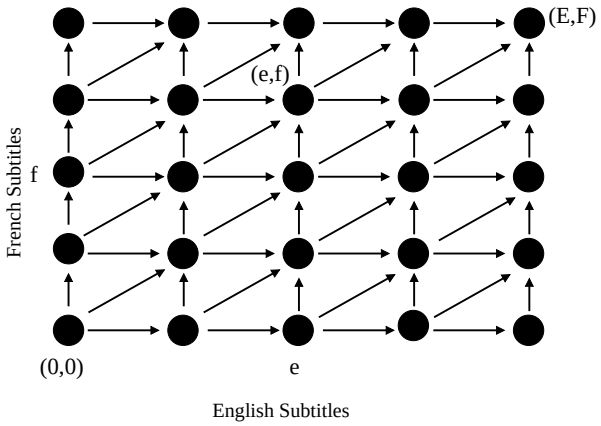

Fig. 2: Dynamic alignment for subtitles

- vertical progress from (e, f) to $(e, f+1)$: the subtitle e corresponds to two consecutive French subtitles
- diagonal shift from (e, f) to $(e+1, f+1)$: the shift towards $(e+1, f+1)$ means that $e+1, f+1$ are potentially translation one of other.
- horizontal transition from (e, f) to $(e+1, f)$: the subtitle f matches with two consecutive English subtitles.

For each node (e, f), we define a matching score based on the F-measure (F_M) calculated as follows:

$$S(e,f) = max \begin{cases} S(e,f-1) + \beta_{F_m}(F_M(e,f) + \varepsilon) \\ S(e-1,f-1) + \alpha_{F_m}(F_M(e,f) + \varepsilon) \\ S(e-1,f) + \lambda_{F_m}(F_M(e,f) + \varepsilon) \end{cases}$$

α_{F_m}, β_{F_m} and λ_{F_m} are tuned to make the alignment as efficient as possible. These coefficients depend on the value of F_M (see section 2.4 for more details). One can notice that the previous formula uses a smoothed F-measure to prevent from a null value. F_M is calculated as follows:

$$F_M(e,f) = 2 \times \frac{R(e,f) \times P(e,f)}{R(e,f) + P(e,f)} \quad (1)$$

e is an English subtitle made up of the words $e_1 e_2 \dots e_{|e|}$ and f is a French subtitle $f_1 f_2 \dots f_{|f|}$. $|e|$ and $|f|$ are the sizes of respectively the English subtitle and the French subtitle. The recall R and the precision P are defined by:

$$R(e,f) = \frac{match(e, tr(f))}{|e|} \qquad P(e,f) = \frac{match(e, tr(f))}{|f|} \quad (2)$$

$tr(f)$ is the list of possible translations for each word in f. $tr(f)$ is obtained by using a French-English dictionary. The function *match* is defined by:

$$match(e, tr(f)) = \sum_{i=1}^{|e|} \sum_{j=1}^{|f|} \delta(e_i, tr(f_j)) \quad (3)$$

$tr(f_j)$ is the set of possible translations of the word f_j, given by the French-English dictionary. For example $tr(\text{fille}) = \{\text{girl}, \text{daughter}\}$. $\delta(e_i, tr(f_j))$ is equal to 1 if the word e_i is in the set $tr(f_j)$ and if e_i does not already match with a previous French word, e.g:

$$\delta(e_i, tr(f_j)) = 1 \Leftrightarrow e_i \in tr(f_j) \text{ and } \forall k < j, e_i \notin tr(f_k) \quad (4)$$

In other cases, $\delta(e_i, tr(f_j))$ is set to 0. The second term of the condition allows to impose that an English word can not match with several occurrences in a French subtitle (as in 'you' in Fig. 3). An example of matching is given in Fig. 3.

To make the matching more accurate, we decided to enhance the *match* function when an orthographic form occurs in both English and French subtitles. This makes, for instance proper names matching without introducing them into the dictionary.

2.4 Test Corpora and protocol

Tests have been conducted on a French-English corpus made up of 40 subtitles files (43013 English subtitles and 42306 French subtitles)[3]. From each movie, we randomly extracted a subset of English and their French corresponding subtitles. This leads to 1353 English subtitles (corpus T_E), and 1334 subtitles in French (corpus T_F). We aligned by hand the selected subtitles. This leads to 1364 (#A) pairs of subtitles (set A) which constitute our reference corpus. We have more pairs of subtitles than the number of subtitles because several consecutive French subtitles may be aligned with several consecutive English subtitles (because of differences in segmentations strategies (Lavecchia et al., 2007)): this 'phrase' alignment leads to several pairs. We used a French-English

[3] extracted from the web-site *http://divxsubtitles.net*

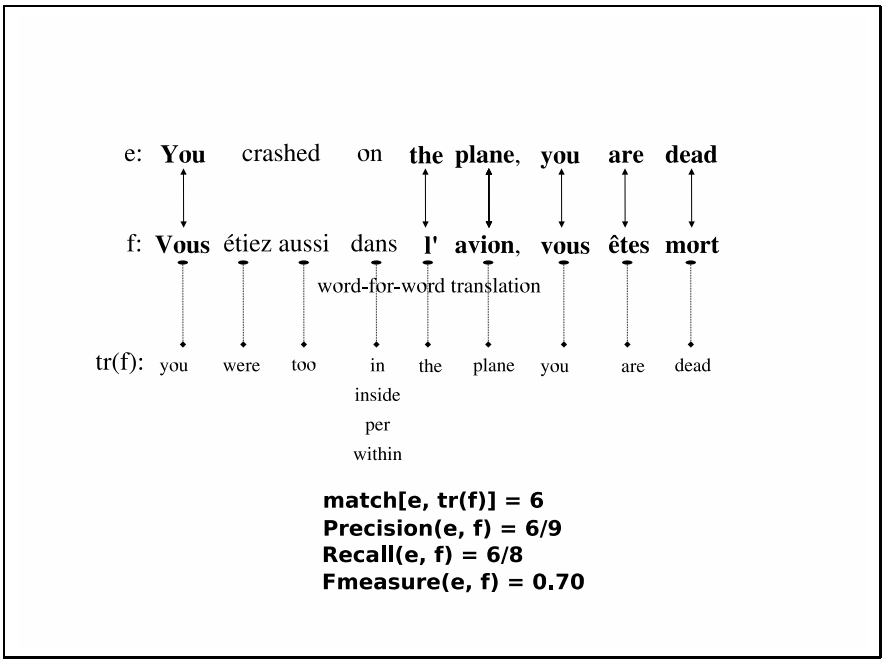

Fig. 3: Illustration of e and f matching

dictionary extracted from the XDXF project[4]. It contains 41398 entries[5]. For the evaluation, we conducted the following procedure:

1. Removing from T_E and T_F subtitles describing events.
2. Alignment of English and French corpora by using DTW based on F-measure.
3. Deletion of the unuseful pairs: each matching pair for which the F-measure is zero is removed. We recall that if the F-measure is zero, then we hypothesize that the subtitles do not match.
4. Comparison with the reference pairs A

2.5 Evaluation

To study the effect of α_{F_M} on the efficiency of our method, we tried several values. In the following experiment, α_{F_M} varies from 1 (the diagonal is not favored) to 100 and β_{F_M} and λ_{F_M} are set to 1. Results in terms of recall, precision and F-measure are presented in Table 1. $\#Tot$. is the number of retrieved pairs. $\#C$ is the number of correct alignments e.g. the number of pairs present in A. $\#I$ indicates the wrong identified pairs. Precision

[4] http://xdxf.revdanica.com/
[5] Archive filename: comn_sdict05_French-English.tar.bz2

and recall are defined by:

$$Precision = \frac{\#C}{\#Tot} \quad Recall = \frac{\#C}{\#A} \tag{5}$$

Table 1: Performance depending on α_{F_M} parameter

α_{F_M}	$\#C$	$\#I$	$\#Tot.$	Rec.	Prec.	Fm.		α_{F_M}	$\#C$	$\#I$	$\#Tot.$	Rec.	Prec.	Fm.
1	1063	842	1905	0.779	0.558	0.650		7	1119	97	1216	0.820	0.920	0.867
2	1124	213	1337	0.824	0.841	0.832		8	1118	96	1214	0.820	0.921	0.867
3	1124	114	1238	0.824	0.908	0.864		9	1119	94	1213	0.820	0.923	0.868
4	1121	99	1220	0.822	0.919	0.868		10	1118	94	1212	0.820	0.922	0.868
5	1121	98	1219	0.822	0.920	0.868		20	1116	93	1209	0.818	0.923	0.867
6	1120	97	1217	0.821	0.920	0.868		100	1114	92	1206	0.817	0.923	0.867

The results showed that α_{F_M} parameter has a strong effect on the performance. We can notice that F_M increases with α_{F_M} until 7 and then the value becomes unstable. In order to set definitely the different parameters we have to remind our objective. In fact, we would like to collect as much aligned subtitles pairs as possible without introducing noise. Table 1 shows that this objective is reached when the precision is maximum. In fact, when precision increases, the number of False Positives[6] decreases. Considering this objective, we decided to set α_{F_M} to 9 in the following experiments. This value leads to 82% of recall and only 94 pairs mismatch. Analyzing results showed that the wrong identified pairs have sometimes a high F-measure. This is due to the weight of small words (prepositions, conjunctions, ...). Such words are uniformly present in several subtitles which make the F-measure positive even if the French and English sentences do not match. This is particularly more critical when subtitles are short as illustrated on Table 2.

Table 2: Illustration of mismatching due to small words

	E1 : Wallis **hold** on to this F1 : Wallace **tiens** moi cela	E1 : Wallis hold on **to** this F2 : Ulrich pense **à**
N(e)	5	5
N(f)	4	3
match	1	1
Prec.	$1/4$	$1/3$
Rec.	$1/5$	$1/5$
F_M	0.22	0.23

[6] the number of incorrect alignments

Two potential pairs of alignment get the same F-measure if their constituents have the same length and the same number of matching words. The alignment (E1, F1) is considered correct whereas the second is wrong. Unfortunately, the F-measure refutes this fact. Indeed, the number of words matching in both pairs is the same but the matching in (E1, F2) concerns two small words (language tool word): *"à"* in French and *"to"* in English. It is obviously incongruous to let these small words having an important influence on the alignment decision. We can point out that the proper name Wallace (Wallis) is missing in the dictionary. A better dictionary coverage (including this proper name) would achieve a F-measure of 0.44 and would allow the couple (E1, F1) to be a better alignment. To reduce the impact of tool words we modified the formula 6 as follows:

$$match(e, tr(f)) = \sum_{i=1}^{|e|} \sum_{j=1}^{|f|} \gamma \times \delta(e_i, tr(f_j)) \quad (6)$$

Where γ is smaller than 1 when e_i or f_j are tool words, otherwise γ is set to 1. Results using (6) are presented in Table 3. This table shows that assigning less weights to tool words unfortunately does not improve results. The more the weight decreases, the more F-measure, Recall and Precision fall down. Naturally a subtitle is short (between 7 and 10 words) and furthermore it is formed by several tool words, it is henceforth difficult to do without this small words. By examining the subtitles pairs proposed by the automatic alignment (with $\alpha_{F_M} = 9$), we discover that 182 out of 1119 correct aligned pairs matched only because of tool words. By decreasing their weight in the match function, we decreased also the F-measure. This could explain also the last line of Table 3. When we omitted tool words (γ set to 0) we noticed that the number of proposed pairs felt considerably. We remind that in the procedure of alignment, we remove all the pairs (e, f) for which the F-measure is equal to 0. That is why all the pairs which matched only on tool words disappeared from the alignment, 289 subtitle pairs are concerned by this cut off. With $\gamma = 1$, we obtained a precision of 92.3% This result is competitive in accordance to the state of art of noisy corpus alignment (Singh and Husain, 2005). These results are very confident and can be used in order to constitute automatic aligned corpora. By launching the developed alignment method with $\gamma = 1$ on the total corpus, we detected 4 files among the 40 leading to a very bad alignment. For the

Table 3: Impact of reducing the tool words' weight

γ	*#C*	*#I*	*#Tot*	Rec.	Prec.	F_M	γ	*#C*	*#I*	*#Tot.*	Rec.	Prec.	Fm.
1.0	1119	94	1213	0.820	0.923	0.868	0.4	1056	171	1227	0.774	0.861	0.815
0.9	1097	134	1231	0.804	0.891	0.845	0.3	1044	189	1233	0.765	0.847	0.804
0.8	1097	134	1231	0.804	0.891	0.845	0.2	1040	192	1232	0.762	0.844	0.801
0.7	1097	134	1231	0.804	0.891	0.845	0.1	1039	194	1233	0.762	0.843	0.800
0.6	1097	133	1230	0.804	0.892	0.846	0.0	869	55	951	0.657	0.942	0.774
0.5	1097	133	1230	0.804	0.892	0.846							

following experiments, we decided to discard these 4 files from the corpus. This final corpus contains 32720 subtitle pairs and leads to a precision of 94%.

In this section, we have described a method to align subtitle files. We have evaluated this method by comparing the alignments with a manual reference. In the following, we propose to use this new parallel corpus to estimate the parameters of a subtitling machine translation system.

3 Translation process

The translation process consists in looking for a E^* sentence which is a translation of a given F sentence. This can be done by estimating the probability $P(E|F)$ and by searching E^* such that:

$$E^* = \arg\max_E P(E|F) \tag{7}$$

Using the Bayes rules, Eq. (7) is rewritten as:

$$E^* = \arg\max_E \frac{P(F|E)P(E)}{P(F)} = \arg\max_E P(F|E)P(E) \tag{8}$$

$P(F|E)$ is defined by the translation model, and $P(E)$ by the target language model. The translation model may be estimated by defining the involved parameters and by using an iterative process which uses the parallel corpus in order to estimate the parameters. This approach is the one chosen in the Giza++ tool (Al-Onaizan et al., 1999).

We propose in the following an original method to construct automatically the translation table without any external knowledge. Each couple of words (e, f) is assigned a probability calculated from inter-lingual triggers. In the following, we describe the idea of inter-lingual triggers, the way to exploit and to make good use of them in order to produce a bilingual dictionary and the translation probabilities $P(e|f)$. Finally, our translation table is evaluated by comparing it to the one achieved by Giza++ (Al-Onaizan et al., 1999) in an entire translation decoding process supplied by Pharaoh (Koehn, 2004).

3.1 A Brief Remind of Triggers

The concept of triggers has been largely used in statistical language modeling. Triggers improve and generalize the Cache model (Kuhn and DeMori, 1990). The Cache model enhances the probability of a word w_i when it occurs in its left context. A trigger model goes further and enhances the probability of a list of words which are correlated to w_i (Tillmann and Ney, 1996). Triggers between two events x and y are determined by computing Mutual Information given by:

$$MI(x,y) = P(x,y) \times log(\frac{P(x,y)}{P(x) \times P(y)}) \tag{9}$$

In statistical language modeling, an event x is the occurrence of a word. For each word x, the n best correlated words in terms of Mutual Information are called the triggered

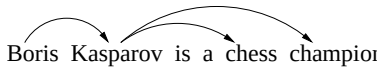

Fig. 4: Examples of triggers

words. x is the triggering word. We call a trigger a set made up of a triggering word and its triggered words. See Fig. 4 for examples of triggers.

Triggers allow to estimate the probabilities of words given a history $P(w|h)$, and are combined with n-grams to achieve a better model (Tillmann and Ney, 1997).

3.2 Inter-lingual triggers

We extended the idea of triggers to inter-lingual triggers. A inter-lingual trigger is henceforth a set composed of a word e in a source language, and its best correlated words in a target language $f_1, f_2, \ldots, f_n$. (see Fig. 5 for examples of inter-lingual triggers). Inter-lingual triggers are determined according to the following formula:

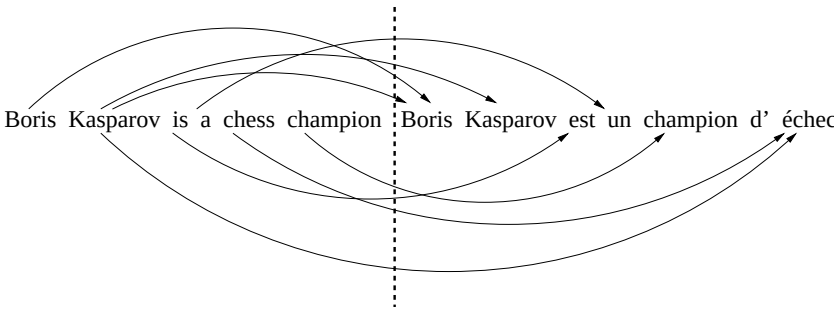

Fig. 5: Examples of inter-lingual triggers

$$MI(f,e) = P(f,e) * log(\frac{P(f,e)}{P(f) * P(e)}) \tag{10}$$

where f (respectively e) is a French (respectively English) word, and, $P(e)$, $P(f)$ and $P(f,e)$ are defined as follows:

$$P(X) = \frac{N(X)}{|Corpus|} \qquad P(f,e) = \frac{N(f,e)}{|Corpus|} \tag{11}$$

where:

$N(X)$ is the number of sentences where X occurs

$N(e, f)$ is the number of sentence pairs where e and f co-occur
$|Corpus|$ is the number of sentence pairs in the training corpus.

For each source word e, we kept the n-best target words $f_1, f_2, \ldots, f_n$ as its triggered words, according to the best MI. This will be written as:

$$Trig(e) \longrightarrow f_1, f_2, \ldots, f_n \tag{12}$$

Inter-lingual triggers have been also used in (Kim and Khudanpur, 2004) to enrich resource deficient languages from those which are considered as potentially important. Our purpose is to use inter-lingual triggers to generate translation tables for machine translation without using any external knowledge. That is why we compute the co-occurrence between e and f inside each sentence pair of the parallel corpus. Clearly, we would like to retrieve the words in a target language $F = f_1, f_2, \ldots, f_n$ which are correlated to a word e in a source language. Among the set F, we hope to find a subset T which is made up only by the translations of e. We compute inter-lingual triggers on a subset of the subtitle corpus described in section 2 (training corpus statistics are given in Table 4).

Table 4: Training Corpus statistics

		French	**English**
Training	Subtitles	27523	
	Words	191185	205785
	Vocabulary	14655	11718
	Singletons	7066	5400

Table 5 illustrates some examples of the obtained English-French triggers, whereas Table 6 gives some French-English triggers. The third column indicates the Mutual Information associated to each couple (trigger and triggered words).

A qualitative analysis showed that our method leads to remarkable inter-lingual triggers where the triggered words could be considered as potential translations of the trigger or very close in terms of meaning.

3.3 Using inter-lingual triggers for building a bilingual dictionary

Our first goal is to provide automatically a bilingual dictionary in multiple languages from inter-lingual triggers. The translations of a source word e are obtained by selecting all the target triggered words $f_1, f_2, \ldots, f_n$ which trigger the source word e as illustrated in Fig. 6. Namely, an entry in a dictionary D is defined as:

$$e: f_1, f_2, \ldots, f_n \in D \Leftrightarrow \forall j \in [1..n], e \in Trig(f_j) \text{ and } f_j \in Trig(e) \tag{13}$$

In this way, we can build French-English and English-French dictionaries. Table 7 gives a view of the obtained French-English dictionary (see section 4.1 for an evaluation of

Table 5: Examples of French words triggered by English words

English trigger word	**French triggered word**	$MI \times 10^{-4}$
Apple	pomme	9.32
	sorbet	3.21
	fruits	2.81
Soldier	soldat	11.93
	combattez	2.88
	soldats	2.17
Chocolate	chocolat	64.06
	chocolaté	7.04
	tablette	6.49
Ladies	mesdames	23.13
	messieurs	20.27
	dames	4.92
Job	boulot	35.79
	travail	34.02
	emploi	8.08

Table 6: Examples of English words triggered by French words

French trigger word	**English triggered word**	$MI \times 10^{-4}$
Pomme	apple	9.32
	baked	2.96
	beef	2.96
Soldat	soldier	11.93
	discovers	3.06
	fighting	1.96
Chocolat	chocolate	64.06
	candy	7.23
	eat	2.81
Mesdames	ladies	23.13
	gentlemen	15.78
	belt	4.24
Travail	job	34.02
	work	30.88
	done	7.58

this dictionary). We called the French-English dictionary Trig-Dic-Reverse because of the membership propriety applied in both English-French and French-English directions.

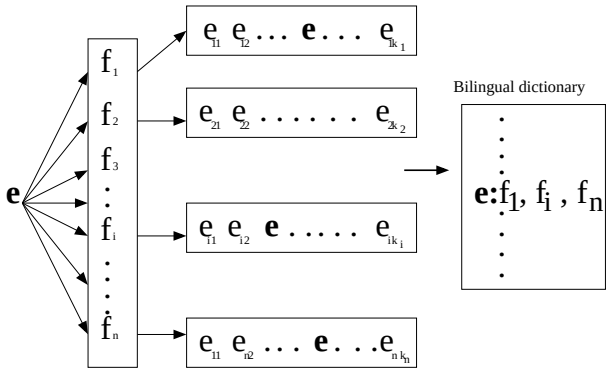

Fig. 6: Illustration of dictionary designing

Table 7: A selection of few entries of French-English dictionary

French word	**Potential translations**		
Demande	ask	wonder	request
Jamais	never	ever	seen
Mort	dead	death	died
Maisons	houses	homes	buy
Pomme	apple	baked	beef

3.4 Using inter-lingual triggers to estimate a translation table

To achieve a translation table using inter-lingual triggers, we assign to each potential word's translation a probability calculated from *MI* such as:

$$\forall f, e_i \in PT(f) \qquad P(e_i/f) = \frac{MI(e_i, f)}{\sum_{e \in PT(f)} MI(e, f)} \tag{14}$$

where $PT(f)$ is the potential translations of f. We estimate several different translation tables:

- The translation table Trig-Dic is directly calculated from Trig-Dic-Reverse with $n = 10$ generated in section 3.3. Each potential translation respects the constraint (13).
- In the translation table Trig-5, we kept as potential translations of a French word f its 5 best triggered words without applying (13).
- In the translation table Trig-10, we kept as potential translations of a French word f its 10 best triggered words without applying (13).
- In the translation table Trig-20, we kept as potential translations of a French word f its 20 best triggered words without applying (13).

4 Evaluation and Experiments

In this section, we first evaluate our dictionary Trig-Dic-Reverse (see section 3.3) by comparing it with two existing dictionaries. Then, we evaluate the different translation tables produced in section 3.4 by integrating them into an entire decoding process supplied by the Pharaoh decoder (Koehn, 2004).

4.1 Evaluation of our French-English dictionary

To evaluate the pertinence of our dictionary (Trig-Dic-Reverse with $n = 5$), we compared it with two dictionaries: one distributed by ELRA[7] and a free downloaded one[8]. The comparison is only done on the French-English side. To make the evaluation relevant, we compare only words which exist in Trig-Dic-Reverse and in the two other dictionaries. Our dictionary share 6098 words with the ELRA dictionary and 6228 with the Internet dictionary. The evaluation in terms of recall is presented in Table 8.

Table 8: Results in terms of recall

	Rank1	**Rank5**
ELRA	16.04	73.74
Internet	13.08	71.11

The results show that if we consider only the translation given in first position, the recall is 16.04%, and if we consider the results without taking care about the rank, the recall reaches 73.74% with ELRA dictionary reference and 71.11% with the Internet dictionary reference. If we consider the ELRA dictionary as a reference, we can say that our algorithm finds out the pertinent translation of a word in 74% of cases. In a first analysis, we can consider that our algorithm has a failure rate of 36%. A deeper analysis contradicts this assertion. In fact, the failure rate can be explained as follows:

- Into Trig-Dic-Reverse we kept only the first five best translations.
- When a potential translation in Trig-Dic-Reverse does not exist in the ELRA dictionary, we notice that frequently the one we propose is correct and sometimes is very close to the meaning.
- In some cases, the translation proposed by ELRA is less commonly used than ours as shown in Table 9. Then even if the translation we propose is correct, it is not counted as correct.

To sum up we can say that the results obtained are very interesting and the recall is probably better than 74%. We have to compare Trig-Dic-Reverse to a better reference (a hand-constructed one) to have a precise evaluation.

[7] M0033-3 SCI-FRAN-EURADIC which contains 70832 entries

[8] http://xdxf.revdanica.com/down/index.php which contains 41398 entries

Table 9: Comparison between ELRA and Trig-Dic-Reverse dictionaries

Word	**ELRA**	**Trig-Dic-Reverse**
chevaux	horseflesh	horses, breed, turbo
chimère	bubble	chimera, monster, virus

4.2 Translation decoding with inter-lingual triggers

In order to evaluate the real contribution of our method, we have to integrate the retrieved translation tables into an entire decoding translation process supplied by Pharaoh[9] (Koehn, 2004). In a first experiment, we use the Trig-Dic translation table generated in section 3.4: each word of source and target language gets 10 potential translations. For each potential translation a probability based on *MI* is associated. The translation probability for other vocabulary words is set to 0. The decoding process has been conducted on a development and a test corpus (Table 10).

Table 10: Development and test Corpus statistics

		French	**English**
Dev	Subtitles	1959	
	Words	13598	14739
Test	Subtitles	3858	
	Words	22195	24729

Table 11: Evaluation of automatic translations using Bleu

Method	Giza++	Trig-5	Trig-10	Trig-20	Trig-Dic
Bleu	0.121	0.113	0.114	0.072	0.113

Translation results in terms of Bleu (Papineni and al., 2001) are given in Table 11. The performance is compared to the one obtained with Giza++ dictionary using the IBM Model 3 (Brown and al., 1993). Table 11 shows that using the 20 best triggers leads to less powerful results. This is probably due to the nature of subtitles, they are in fact short. Consequently, it is too difficult to find 20 target words which are correlated to a source word within a subtitle. The decoding results based on our method is similar to

[9] The target language model is a trigram model (Good-Turing smoothing, cutoff set to 7 for bigrams and trigrams). The decoding weights are set to: 1 for language model, 1 for translation model, 0 for word penalty, and 1 for distortion model. Decoding is with reordering.

the one achieved by Giza++ and the best result is obtained with Trig-10. Furthermore, Giza++ trains IBM models 1, 2 and 3 in several iterations to outperform slightly our model. Table 12 shows that until the fifth iteration of IBM2, Giza++ does not outperform Trig-10 whereas Trig-10 needs only one iteration.

Table 12: Bleu Evolution

Model-Iteration	M1-it1	M1-it5	M2-it1	M2-it5	M3-it5	Trig-10
Bleu	0.075	0.096	0.097	0.099	0.121	0.114

To improve results, we optimize the Pharaoh parameters for all the decoders (Trig-Dic-10, Trig-Dic and Giza++). Table 13 presents the performance with tuned parameters. In this table, tm, lm, dm and w are respectively the decoding weights for the translation model, the language model, the distortion model and the word penalty. With optimal parameters of Trig-10, we outperform Giza++ by 10,9% whereas with the optimal parameters of Giza++, our model is 6% worse. This orientation is maintained and emphasized

Table 13: Optimization of Pharaoh parameters for Trig-10, Trig-Dic and Giza++ decoders

	tm	lm	dm	w	Bleu-Trig	Bleu-Giza++
Trig-10	0.9	0.4	0.4	0	0.1227	0.1105
Trig-Dic	0.9	0.4	0.3	-2	0.1184	0.1094
Giza++	0.7	0.6	0.5	-2	0.1157	0.1220

on the test corpus (see Table 14). In an optimal use of Trig-10, this one outperforms Giza++ by 8,1% whereas Giza++, in an optimal use, does better than Trig-10 by only 1,2%. Furthermore, the best Bleu score for the translations obtained with the Giza++ translation table is 0.1176 whereas the best Bleu score for the translations obtained with the Trig-10 translation table is 0.1202. In other terms, our method outperforms slightly IBM model 3 by 2.2%.

Table 14: Decoder evaluation with optimal parameters for Pharaoh on the test corpus

tm	lm	dm	w	Trig-10	Giza++
0.9	0.4	0.4	0	**0.1202**	0.1119
0.7	0.6	0.5	-2	0.1161	0.1176

5 Conclusion and future work

In this paper, we presented a full process of translation, starting at the alignment process and ending at translation decoding. We built a movie subtitle parallel corpora of 32720 aligned pairs with a precision of 94% compared to a manual alignment. Then, this material has been used to construct a dictionary based on inter-lingual triggers. For each word (French or English) a list of its target corresponding triggers has been proposed. An entry of a bilingual dictionary is made up of a source word and its best target triggers. The obtained dictionary is relevant and first results in an entire decoding process showed that on a test corpus with an optimal set of parameters of Pharaoh, our method outperforms Giza++ by 8,1%. With the optimal parameters of Giza++, this one over-pass our method by only 1,2% . Furthermore, the best Bleu score for the translations obtained with the translation table Trig-10 is 2.2% better than the best Bleu score of Giza++ translation table. Our results are very encouraging and efforts are done in order to improve our model by using phrases in the translation decoding process. The idea of using cross lingual triggers seems to be very attractive, it can be used in several areas in machine translation. For instance, they could be used as a confident measure. Several other utilizations of this method have been imaginated and are under-work in our research group.

Bibliography

[Al-Onaizan et al., 1999]Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, I., Och, F., Purdy, D., Smith, N., and Yarowsky, D. (1999). Statistical machine translation. In *Final Report, JHU Workshop*.

[Brown and al., 1993]Brown, P. F. and al. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:263–311.

[Brown et al., 1991]Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Meeting of the Association for Computational Linguistics*, pages 169–176.

[Gale and Church, 1991]Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184.

[Kim and Khudanpur, 2004]Kim, W. and Khudanpur, S. (2004). Lexical triggers and latent semantic analysis for cross-lingual language model adaptation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(2):94–112.

[Koehn, 2004]Koehn, P. (2004). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *6th Conference Of The Association For Machine Translation In The Americas*, pages 115–224, Washington, DC, USA.

[Koehn, 2005]Koehn, P. (2005). Europarl: A multilingual corpus for evaluation of machine translation. In *MT Summit*, Thailand.

[Kuhn and DeMori, 1990]Kuhn, R. and DeMori, R. (1990). A cache-based natural language model for speech recognition. *IEEE Trans. PAMI*, 12(6):570–582.

[Lavecchia et al., 2007]Lavecchia, C., Smaili, K., and Langlois, D. (2007). Building parallel corpora from movies. In *Proceedings of The 5th International Workshop on Natural Language Processing and Cognitive Science*, Funchal, Madeira - Portugal.

[Melamed, 1996]Melamed, I. D. (1996). A geometric approach to mapping bitext correspondence. In Brill, E. and Church, K., editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1–12. Association for Computational Linguistics, Somerset, New Jersey.

[Moore, 2002]Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the Association for Machine Translation in the Americas Conference*, pages 135–144.

[Papineni and al., 2001]Papineni, K. and al. (2001). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual of the Association for Computational linguistics*, pages 311–318, Philadelphia, USA.

[Singh and Husain, 2005]Singh, A. K. and Husain, S. (2005). Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and using Parallel texts*, pages 99–106.

[Tillmann and Ney, 1996]Tillmann, C. and Ney, H. (1996). *Selection criteria for word trigger pairs in language modeling*, pages 98–106. Lecture Notes in Artificial Intelligence 1147, Springer Verlag.

[Tillmann and Ney, 1997]Tillmann, C. and Ney, H. (1997). Word trigger and the EM algorithm. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 117–124, Madrid, Spain.

[Vandeghinste and Sang, 2004]Vandeghinste, V. and Sang, E. K. (2004). Using a parallel transcript/subtitle corpus for sentence compression. In *LREC*, Lisbon, Portugal.